DOI 10.1007/s10586-015-0523-z

Robust audio fingerprinting using peak-pair-based hash of non-repeating foreground audio in a real environment

Hyoung-Gook Kim[1] · Hye-Seung Cho[2] · Jin Young Kim[3]

Received: 30 October 2015 / Revised: 8 December 2015 / Accepted: 19 December 2015 / Published online: 2 January 2016

Abstract In this paper, we propose a high-performance audio fingerprinting system used in real-world query-by-example applications for acoustic audio-based content identification, especially for use in heterogeneous portable consumer devices or on-line audio distributed system. In the proposed method, audio fingerprints are generated using a modulated complex lapped transform-based non-repeating foreground audio extraction and an adaptive thresholding method for prominent peak detection. Effective matching is performed using a robust peak-pair-based hash function of non-repeating foreground audio to protect against noise, echo, artifacts from pitch-shifting, time-stretching, resampling, equalization, or compression. Experimental results confirm that the proposed method is quite robust in various distorted conditions and achieves preliminarily promising accuracy results.

Keywords Audio fingerprinting · Modulated complex lapped transform · Peak detection · Robust hash function

✉ Hyoung-Gook Kim
hkim@kw.ac.kr

Hye-Seung Cho
hye_seung401@kw.ac.kr

Jin Young Kim
beyondi@jnu.ac.kr

[1] Department of Electronics Convergence Engineering, Kwangwoon University, 20 Gwangun-ro, Nowon-gu, Seoul, Korea

[2] Department of Radio Sciences and Engineering, Kwangwoon University, 20 Gwangun-ro, Nowon-gu, Seoul, Korea

[3] Department of Electronics and Computer Engineering, Chonnam National University, 77 Yongbong-ro, Buk-gu, Gwangju, Korea

1 Introduction

Audio fingerprinting is a technology to identify short, unknown audio clips in a labeled audio database based on a compact set of audio fingerprints in a fast and reliable way [1]. It can be used in many applications or services for mobile consumer devices or distributed systems, such as music retrieval [2], duplicate song detection [3], advertisement tracking [4], broadcast monitoring, copyright detection, filtering for file sharing, and automatic audio-based library content organization [5].

For reliable identification in real-world applications, three important requirements [6] should be satisfied: (i) it should be able to identify corrupted query audio clips in a given music collection despite the degradations by different noise and artifacts including pitch-shifting, time-stretching, equalization, or dynamics compression; (ii) it should be able to identify audio clips that are only a few seconds long; and (iii) it should be able to perform the task for both fingerprint generation and searching in a large database in a computationally efficient manner.

To satisfy these practical requirements for a successful audio fingerprinting system, a number of audio fingerprinting system [7] have been developed in recent years. The audio fingerprinting system proposed by Haitsma et al. [5] is one of the most widely used systems. It presents an audio fingerprint based on quantized energy changes across two-dimensional time-frequency space. However, this technique has drawbacks that the amount of information is relatively large and it has poor performance in low signal-to-noise ratio (SNR) conditions [8]. The approach taken by Baluja et al. [9] proposes a method based on sparse wavelet representation of overlapping spectrogram images and a min-hash technique. Their approach is computationally very expensive and results in a high number of bits per fingerprint [10], however it provides a high identification rate.

The system developed by Wang [11] is effective both in calculating the fingerprints and in searching for the best match in a large database using three important properties: spectral peak, a uniform selection of spectral peaks and landmark hashes. For this reason, Wang's method is the most commercially successful and wide spread work, and various methods based on Wang's idea have been proposed to achieve high identification accuracy in real environments. Pan et al. [8] used a local energy centroid as a method of extracting an audio fingerprint, whereas a prominent peak detection based on modulated complex lapped transform was presented by Kim et al. [12] for audio fingerprinting. Fenet et al. [13] proposed a hashing technique coupled with a constant Q transforms-based fingerprint that includes robustness to pitch-shifting. An approach using spectrogram masking around spectral points was proposed by Anguera [10].

We think that three properties of Wang's method can be improved: (i) spectral peaks (defined as a local maximum in the logarithmic magnitude spectrum) on short-time Fourier transform (STFT) are highly characteristic, reproducible, and robust against distortions by the environmental noise or equalization. However, the percentage of surviving peaks (that exist in both pure and noisy audio clips) is reduced when short query audio clips are degraded by environmental noise. To detect sufficient and distinct spectral peaks for audio comparison in spite of short audio clips, the STFT should be replaced by another efficient transform method. And an efficient background noise reduction method should be needed; (ii) a uniform selection of spectral peaks using dynamic threshold methods reduces the complex spectrogram to a low-dimensional sparse representation of the audio signal. However, a strong peak-picking thresholding method hinders the extraction of the spectral peaks for providing effective matching between the audio query clip and audio collection. It should be able to extract important spectral peaks using an adaptive peak-picking threshold; and (iii) spectral peak locations (defined as a combination of both the frequency values and the time difference between the peaks) are used as effective fingerprint hashes. They try to efficiently reduce the retrieval time using indexing techniques. The hashes consisting of a start frequency, frequency differences, and time differences between the peaks in the target zone, are not effective against pitch-shifting and time-stretching. To overcome these difficulties, robust hashes are needed.

In this paper, a novel audio fingerprinting method using peak-pair-based hash of non-repeating foreground audio is proposed to improve the robustness of audio fingerprinting in real environments.

The major contribution of the proposed method are as follows: (i) modulated complex lapped transform (MCLT) [14] are used to extract the majority of the sound's local spectral peaks more effectively than STFT in spite of short audio clips; (ii) to achieve high identification accuracy against various noise and echo conditions, non-repeating foreground audio is extracted from repeating background audio and used in the generation of peak-pair audio fingerprints; (iii) to obtain salient peak pairs against different types of distortions and at different distances from the audio source, an adaptive thresholding method based on peak-picking update of the non-repeating audio pattern is applied; and (iv) using novel fingerprint hashes, the proposed algorithm improves the robustness of the audio fingerprinting against equalization, compression, pitch-shifting, and time-stretching.

This paper is organized as follows. Section 2 describes the proposed method. Section 3 discusses the experimental results. Finally, the conclusion is presented in Sect. 4.

2 Proposed robust audio fingerprinting method

The two key components of the fingerprinting system are the fingerprint extraction and the fingerprint matching. A fingerprint is extracted from a short audio clip captured by a fingerprint client such as a portable consumer device and submitted to the fingerprint server. The extracted fingerprint is then used to query the fingerprint database at the server and is compared with the stored fingerprints.

In Wang's method, each audio track is analyzed using the STFT to extract local spectral peaks that have amplitude that is larger than the peaks in a surrounding area. The time and frequency distances between pairs of these peaks are encoded into a landmark hash that represents the audio. To identify a query, it is similarly converted into landmarks. Then, the database is queried to find all the reference tracks that share landmarks with the queries, and the relative time differences between where they occur in the query and where they occur in the reference tracks.

For robust fingerprint extraction against noise and various types of distortion, we propose using a peak pair-based hash of non-repeating MCLT foreground audio based on Wang's idea.

Figure 1 illustrates the proposed robust audio fingerprint extraction method, which is composed of six main blocks.

First, a stereo audio signal is captured by the user's smartphone and is converted into mono, downsampled to 16 kHz. And the converted audio signal is segmented into overlapping frames by the application of a Hanning window function (each of which contains 512 overlapped samples). Then, the MCLT is applied to each frame (1024 samples) to find the spectral peaks. After separating each repeating background and each non-repeating foreground audio part from the MCLT spectrogram, a logarithmic operation of each non-repeating foreground is performed. The normalized logarithmic non-repeating foreground MCLT spectrogram

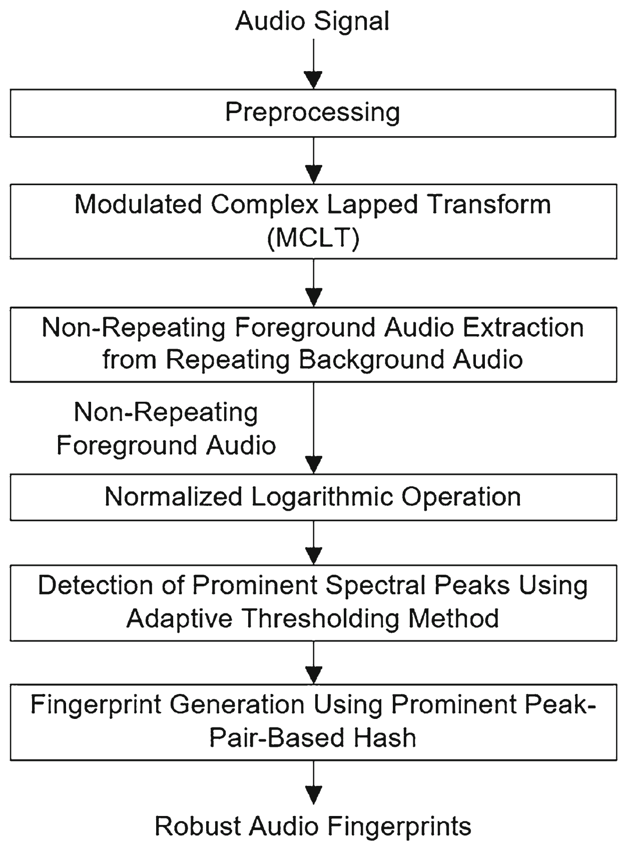
Audio Signal
Preprocessing
Modulated Complex Lapped Transform (MCLT)
Non-Repeating Foreground Audio Extraction from Repeating Background Audio
Non-Repeating Foreground Audio
Normalized Logarithmic Operation
Detection of Prominent Spectral Peaks Using Adaptive Thresholding Method
Fingerprint Generation Using Prominent Peak-Pair-Based Hash
Robust Audio Fingerprints

Fig. 1 Block diagram of robust audio fingerprint extraction

containing high energy is fed into a prominent spectral peak detection step, where the salient peaks are selected by applying an adaptive temporal threshold. In the local target area of the frequency-time plane, nearby salient peaks are combined into a pair or landmark. Three components using peak-pair location is converted into a 32-bit fingerprint hash in consumer devices, and submitted to the fingerprint server for the acoustic audio-based content identification.

2.1 Time-to-modulation complex lapped transform

The audio signal $s(n)$ is divided into a Hanning-windowed overlapping frame and analyzed using MCLT. The MCLT has a complex-valued portion based on a $2N$ point fast Fourier transform (FFT) $U(k,l)$ of $s(n)$ multiplied by a factor $b(k,l)$:

$$S_M(k,l) = |jV(k,l) + V(k+1,l)| \tag{1}$$

using

$$V(k,l) = b(k,l) \cdot U(k,l) \tag{2}$$

$$U(k,l) = \sqrt{\frac{1}{2N}} \sum_{n=0}^{2N-1} s(n+lO)h(n)exp\left(\frac{-j2\pi kn}{N}\right), \tag{3}$$

$$b(k,l) = W_8(2k+1,l) \cdot W_{4N}(k,l), \tag{4}$$

$$W_T(k,l) = exp\left(\frac{-j2\pi k}{T}\right), \tag{5}$$

where k is the frequency bin index, l is the time frame index, h is an analysis window of size N, O is the framing step.

Unlike orthogonal transforms such as FFT, MCLT has significant overlap in its frequency response for the basis functions and provides twice-frequency resolution. Therefore, MCLT has approximate shift invariance properties [14]. The spectral peaks detected by MCLT preserve the majority of the original peaks of the sound more effectively than the spectral peaks based on STFT against different distortions caused by additive noise, additive echo, and coding artifacts; a sufficient number of distinct peak pairs can therefore be identified as coming from the same reference track when short query audio clips are used.

2.2 Extraction of non-repeating foreground audio from repeating background audio

The MCLT spectrogram $S_M(k,l)$ is composed of structures, where a repeating background audio part is superimposed with a varying non-repeating foreground audio part. To separate each repeating background from each non-repeating foreground in the MCLT spectrogram, we modified the repeating pattern extraction technique [15,16].

The method seeks to identify repeating/similar elements in the MCLT spectrogram by using the cosine similarity $C(l_a,l_b)$ between transposed $S_M(k,l_a)$ and $S_M(k,l_b)$, after normalization of the columns of $S_M(k,l_a)$ by their Euclidean norm. The calculation of the similarity matrix C is defined as

$$C(l_a,l_b) = \frac{\sum_{k=1}^{K} S_M(k,l_a)S_M(k,l_b)}{\sqrt{\sum_{k=1}^{K} {S^2}_M(k,l_a)}\sqrt{\sum_{k=1}^{K} {S^2}_M(k,l_b)}} \tag{6}$$

where each point (l_a,l_b) in C measures the cosine similarity between the time frames l_a and l_b of S_M.

The searched repeating elements are applied for building a model of the repeating background using a two-dimensional median filter in the MCLT spectrogram. The repeating background model is then used to construct a soft time-frequency mask to extract the non-repeating foreground $F(k,l)$ from the repeating background $B(k,l)$. The resulting

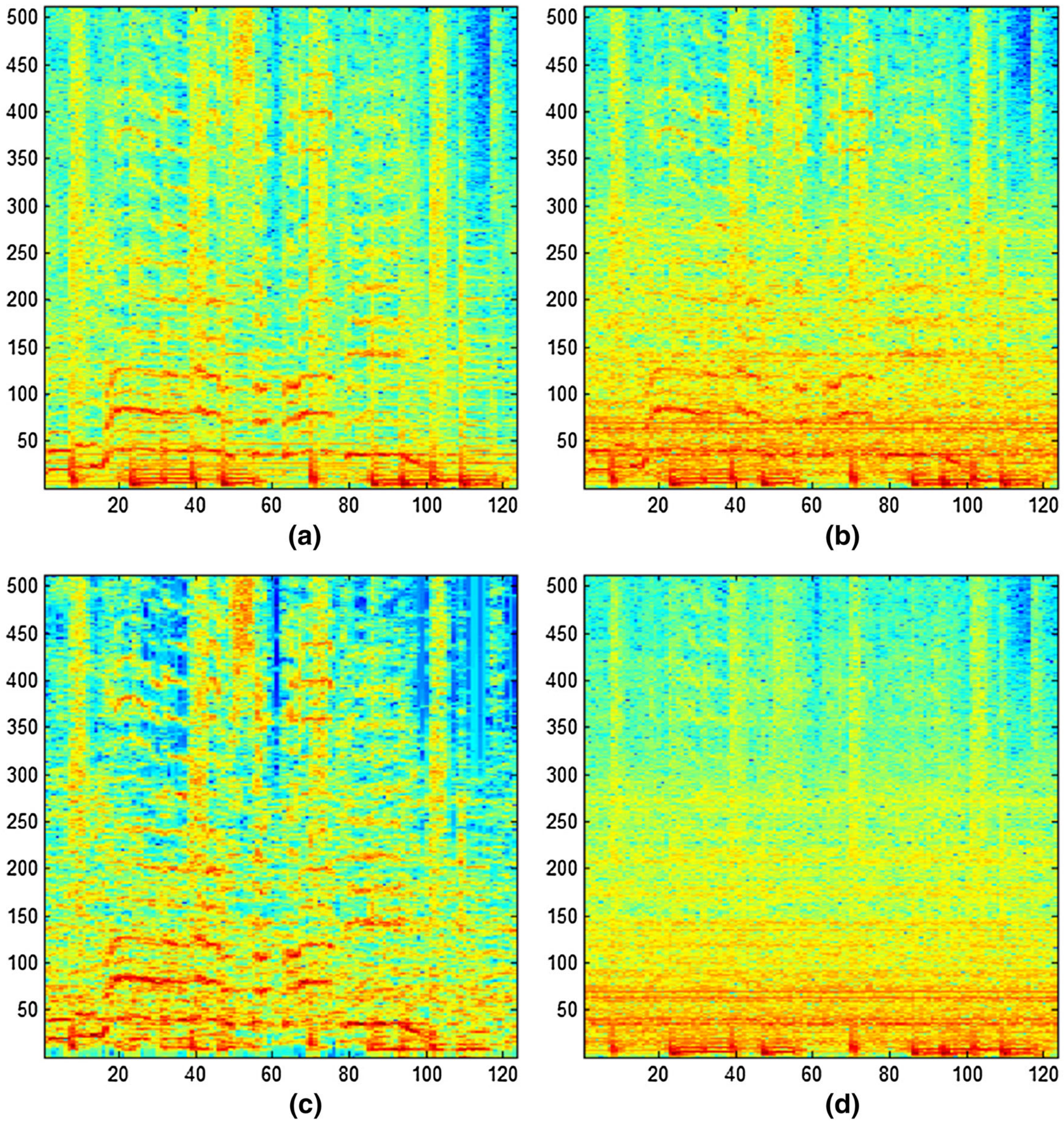

Fig. 2 Spectrogram of the clean music (**a**), spectrogram of the music mixed with SNR 6 dB (**b**), spectrogram of the foreground audio extracted from the music with SNR 6 dB (**c**) and spectrogram of the background audio extracted from the music with SNR 6 dB (**d**)

non-repeating foreground $F(k, l)$ preserves harmonic spectral peaks with local high energy, whereas the background area contains repeating parts with local low energy. For this, time distance between two consecutive repeating frames is selected as 0.3 s. The detailed algorithm is shown in the repeating pattern extraction technique [15,16].

Figure 2 shows examples of the extraction of non-repeating foreground from repeating background in the MCLT spectrogram.

2.3 Detection of prominent foreground spectral peaks

Spectral peaks are highly characteristic, reproducible, and robust against many, even significant distortions of the signal. Therefore, we apply the non-repeating foreground audio to the idea of Wang's algorithm to extract local spectral peaks for their robustness against noise and different types of distortions.

First, a logarithmic operation of each non-repeating foreground $F(k, l)$ is performed by:

$$F_{log}(k, l) = log_{10}(F(k, l)). \tag{7}$$

From the logarithmic non-repeating foreground spectrogram, the mean of the logarithmic non-repeating foreground $F_{log}(k, l)$ is estimated and subtracted in every frame l to generate the normalized logarithmic non-repeating foreground $F_N(k, l)$.

To achieve a small set of time-frequency points or peaks, we apply a prominent peak extraction method to select only those peaks whose energy stays above an adaptive temporal threshold. The prominent peak extraction based on adaptive peak-picking threshold is divided into five steps, as follows:

1. Initial threshold estimation: peak values of each frequency bin in the first frame ($l = 0$) of the normalized logarithmic non-repeating foreground spectrogram (NLNFS) are linear interpolated across frequency bins and used as an initial temporal threshold for peak selection.
2. Prominent peak selection for the second frame: for the second frame ($l = 1$) of the NLNFS, all peaks above the initial temporal threshold are selected as high peaks. Thereafter, the temporal threshold is dynamically updated by raising the initial temporal threshold for the third frame.
3. Prominent peak selection for the further frame: for the next frame ($l = 2$), all peaks higher than the updated smoothing threshold $Th(k,l)$ are stored in a set of (k,l) tuples named salient peaks (SP). If the SP is selected among the $F_N(k,l)$, the SP is represented as $P(k,l)$:

$$F_N(k,l) = \begin{cases} SP, & \text{if } F_N(k,l) > Th(k,l) \\ non - SP, & \text{otherwise} \end{cases} \tag{8}$$

using

$$\begin{aligned} Th(k,l) &= \alpha \cdot Th_{UP}(k,l) + (1-\alpha)\cdot \\ &\quad P(k,l-1) * exp\left(-\frac{m^2}{2\sigma^2}\right) \end{aligned} \tag{9}$$

where m is the distance (in frames) between the previous peak and the considered peak, α, σ represent the smoothing parameter and the parameter used to set the threshold's width, respectively.

4. Adaptive smoothing threshold update: If the SP higher than the smoothing threshold $Th(k,l)$ is searched, the smoothing threshold is adaptively updated by raising the previous temporal threshold with the convolution of all new SP with a point spreading function exp(●):

$$\begin{aligned} Th_{UP}(k,l) &= max\,(Th(k,l-1), P(k,l)* \\ &\quad exp\left(-\frac{m^2}{2\sigma^2}\right)); \end{aligned} \tag{10}$$

5. Steps 3–4 are repeated for $l = l + 1$ until all frames are processed.

The applied adaptive peak-picking thresholding method helps to extract more salient and distinct peak pairs for comparing the query fingerprint with the original fingerprints in spite of various distortions without backward peak selection step that is used in Wang's approach.

2.4 Fingerprint hashes using peak pairs of non-repeating foreground audio

In Wang's approach, fingerprint hashes are generated from combinations of peak pairs named landmark points. Each landmark is selected as an anchor point and paired with nearby landmarks from within a target zone. The target zones include a fixed number of points so each landmark point generates a fixed number of pairs.

Assuming that $P_A(k_a, l_a)$ is the anchor point and paired with another landmark point $P_B(k_b, l_b)$ its hash was obtained by:

$$hash = \left(k_a, k_a - k_p, (l_a - l_p)\right) = (k_a, \Delta k, \Delta l). \tag{11}$$

All k (in frequency bins) and l (in frames) are integers with a fixed higher bound, so each landmark point generates a fixed number of pairs.

The hash method in Wang's approach has high robustness to noise and facilitates a high identification rate, while scaling to large databases. However, this hash method cannot handle pitch-shifting or time-stretching of the audio signal, because time-stretching and pitch-shifting of the query fragment completely change the hash values.

To deal with this problem, we propose a new hash method as follows:

$$\begin{aligned} hash &= \left(\varepsilon \cdot \tfrac{k_a}{k_p}, \theta \cdot \tfrac{k_a - k_p}{k_a}, (l_a - l_p) + (k_a - k_p)\right) \\ &= \left(\varepsilon \cdot k_R, \theta \cdot \tfrac{\Delta k}{k_a}, \Delta l + \Delta k\right). \end{aligned} \tag{12}$$

The first component k_R is the start frequency-to-end frequency ratio (SER) value of a pair of peaks, while the second component $\Delta k / k_a$ represents the frequency extent-to-start frequency ratio (FSR) value of the pair of peaks. The first and second components of the peak-pair in the NLNFS domain show good robustness to both pitch-shifting and time-stretching as shown in Tables 1 and 2, respectively. However, the ratio-based hash values are distributed in small regions, quite intensively, in many cases. To overcome this problem, we place the hash values in various areas by assigning the weighting factors ε, θ to the ratio-based hash values.

In the Tables 1 and 2, ST, ET, SF, EF, and FD denote the start time, the end time, the start frequency, the end frequency, and the frequency extent of a pair of peaks, respectively.

As the third component, we use the summation ($\Delta k + \Delta l$) of the frequency extent Δk and the time extent Δl of the pair in the NLNFS domain. The reason for this is as follows: Time-stretching refers to the process of the changing speed or duration of an audio signal without affecting its pitch and pitch-shifting. Under time-stretching, the component Δk of the peak-pair is invariant whereas the component Δl is changed in accordance with the time-stretching

Table 1 Hash parameter comparisons for a pitch-shifting example

Pitch-shifting (%)	ST	ET	SF	EF	FD	SER	FSR
+40	102	114	366	291	75	1.25	0.20
+13	102	114	293	233	60	1.25	0.20
0	102	114	259	206	53	1.25	0.20
−11	102	114	231	184	47	1.25	0.20
−30	102	114	183	146	37	1.25	0.20

Table 2 Hash parameter comparisons for a time-stretching example

Time-stretching (%)	ST	ET	SF	EF	FD	SER	FSR
+82	59	64	42	94	52	0.45	1.24
+20	40	43	42	94	52	0.45	1.24
0	34	37	42	94	52	0.45	1.24
−20	28	30	42	94	52	0.45	1.24
−30	24	25	42	94	52	0.45	1.24

ratio. Therefore, the component Δk provides good robust performance under time-stretching. On the other hand, pitch-shifting refers to the process of changing the pitch of an audio signal without affecting its speed. Under pitch-shifted condition, the component Δl provides good robust performance. Thus, the third component is robust to cropping, time-stretching, and pitch-shifting within the defined target zone. However, the identification may be decreased, if time-stretching and pitch-shifting occur over the defined target zone.

2.5 Fingerprint matching

When building the fingerprint database at the server, a database index is created by storing the fingerprint hash and Track ID. In the retrieval processing [9], the effective matching is performed using a robust hash function in the fingerprinting domain, as follows:

1. A query signal is fingerprinted in the user's smartphone and submitted to the server. The query file's hashes are compared against the hashes stored in the database hash table at the server. That is to say, all database hashes $\{h_{DB}\}$ matching the query's set of hashes $\{h_Q\}$ are retrieved first.
2. From the hash values of the matched database hashes, the song $\{S_{ID}\}$, the time extent tR and the frequency extent f_R are conversely detected as the reference. Next, the number of the song $\{S_{ID}\}$ that has the same song ID among the matched hash results is counted, and sorted in ascending order. From the sorted list, only the top 30 % results containing the highest count number of the song $\{S_{ID}\}$ are selected as candidate matching results.
3. The time differences $\{t_R - t_Q\}$ and the frequencies differences $\{f_R - f_Q\}$ between candidate reference fingerprints R and the query fingerprints Q are efficiently computed at the same time. We store these differences in histograms (one histogram per candidate reference). If the analysis frame (query frame) is actually an excerpt of the reference, the histogram shows the lowest value and the count of the same histogram value is maximal. If the files match (a sufficient number of landmarks have been identified as coming from the same reference track, with the same relative timing), matching hashes should occur at similar relative offsets from the beginning of the matching file.
4. The final matching process is divided into three cases according the query file distorted by noise, artifacts from pitch-shifting, time-stretching, echo, resampling, equalization, or compression:

- If the query file is distorted by environmental noise, artifacts from echo, resampling, equalization, or compression, its sum of the time differences and the frequency differences of the query file in the candidate files of the database is within predefined threshold (lowest histogram), and the count number of the founded song S_{ID} has the highest maximum, this reference is considered to match the query frame;
- If the query file is pitch-shifted, the time differences are under the predefined time threshold whereas the frequency differences are above the predefined frequency threshold. In this case, we ignore the histogram of the frequency differences and use only the histogram of the time differences for providing the matching results;
- If the query file is time-shifted, the time differences are above the predefined time threshold whereas the frequency differences are under the predefined frequency threshold. In this case, we ignore the histogram of the time differences and use only the histogram of the frequency differences for providing the matching results.

In addition, a statistical filter to remove continuity and redundancy were applied to the matching process for improving the query and response accuracy from the database.

3 Experiments

3.1 Experiment data

To evaluate the performances of the proposed method, two test database types were used: (i) Set I consists of a database of 9000 songs from various genres such as pop, rock, hip-hop, folk, rap, metal, jazz, blues, opera, dance, and classical; and

(ii) Set II is a database containing 8000 TV advertisements with total time amounting to 1580 h.

All the audio data in a labeled audio database is stored in PCM format using mono, 16-bit depth, and 16 kHz sampling rate converted from audio CDs. To cover a wide variety of robustness requirements for real-world application scenarios, the query sets are created by adding various types of distortions with 10,000 randomly selected queries (5000 queries for songs and 5000 queries for TV advertisements) from two test databases, as follows:

- Noise addition: from clean to SNR 10, 5 and 0 dB were added using different types of noise such as babble noise, car noise, white noise, street noise, train noise, airport noise, and computer fan noise;
- Resampling: downsampling to 8 kHz or 32 kHZ and then upsampling back;
- Equalization: gain −5 and 3 dB from 31 Hz to 16 kHz;
- Echo addition: from 100 to 500 ms, 50 % echo addition;
- MP3 encoding/decoding: 64 kbps;
- Time-stretching: from −20 to +20 %. Only the tempo changes, the time duration remains unaffected;
- Pitch-shifting: from −30 to +30 %. Only the pitch changes, the pitch remains unaffected.

Audio query clips with lengths of 2, 3, 4, and 5 seconds were captured through a built-in fingerprint generation module in a smartphone, which was placed 5 m from a 2.1-channel loudspeaker connected to a TV and a radio. Each query audio sample is played 20 times at randomly set offsets.

3.2 Experiment Results

To evaluate the effectiveness of the proposed method, the following five methods (which are based on Wang's idea) have been modified from the contents of the reference papers and then implemented: (i) method 1 (WL) is an STFT-based peak-pair fingerprint extraction method proposed by Wang [11]. To achieve a good identification rate, we optimized various parameters used in Wang's method; (ii) method 2 (EC) is an audio fingerprint method using sub-fingerprint masking based on local energy centroids [8]. To limit the number of maximum points obtained by the local energy centroid using a weighed window functions, an additional temporal masking threshold was applied; (iii) method 3 (MA) is an audio fingerprint extraction method based on the masked audio spectral keypoints [10]. Parameters for the temporal masking threshold and different mask sizes were tested. Among these, reasonable parameters were finally selected; (iv) method 4 (MF) is a fingerprinting method using prominent peak pair based on MCLT [12]. Prominent peaks are formed into pairs within a target area; (v) method 5 (RP) is a fingerprinting method using a hashing technique coupled with a constant Q transform-based fingerprint [13]. Various values were used to find a rough frequency location for the pair of peaks tested to obtain a reasonable identification rate; and (vi) the proposed method is denoted as PM.

Table 3 Comparative performance for the six schemes with set I

SNR	AIAR (%)					
(dB)	PM	WL	EC	MA	MF	RP
Clean	98.3	93.1	93.5	91.6	96.1	94.1
10	96.5	90.4	92.3	88.3	92.7	93.3
5	94.6	81.5	81.4	76.7	90.5	84.6
0	83.4	68.8	69.2	63.5	78.3	71.4
Total	93.2	83.5	84.1	80.0	89.4	85.9

Accuracy is the foremost requirement in most of audio fingerprinting systems. To evaluate the identification performance of the proposed method, we define a performance measure using identification accuracy as follows:

$$Accuracy = \frac{tp}{tp + fp + fn} \tag{13}$$

where tp (true positives) is the number of case in which the correct reference is identified from the query, fn (false negatives) is the number of cases in which the system fails to return a reference id at all, and fp (false positives) is the number of cases in which the system predicts the wrong reference.

Table 3 shows the averaged identification accuracy results (AIAR) of the six methods under five different noisy environments when a 5-second-long query from set I was used.

The AIAR indicates the percentage of queries which were identified as the reference music or song with the most matched fingerprints. The identification results under the five different noisy environments are averaged for the evaluation.

As depicted in Table 3, the best identification accuracy was 98.3 %, which was obtained by the proposed method, PM. The identification performance is decreased as the level of added noise is raised. According to these results, the PM is sufficiently robust against noise compared with the other methods although the noise level increases. The identification accuracy rates of MF are higher than those of WL, EC, MA, and RP whereas the MA yields the lowest identification accuracy rate.

Table 4 shows the identification performance of the PM scheme for when the query length was changed.

These results in Table 4 show that the performance increases as the length of the query increases. Also, the proposed scheme shows satisfactory performance with 4 and 5-second-long queries, showing the identification accuracy rates above 90 %. Especially, the identification accuracy rates using 2-second-long queries are remarkably decreased com-

Table 4 Performance evaluation according to query length with set I

SNR	AIAR by query length (%)			
(dB)	2 s	3 s	4 s	5 s
Clean	78.1	92.6	96.3	98.3
10	73.4	91.8	95.2	96.5
5	65.6	86.2	91.9	94.6
0	57.4	78.4	80.3	83.4
Total	68.6	87.3	90.9	93.2

Table 5 Comparative performance for the six schemes with set II

SNR	AIAR (%)					
(dB)	PM	WL	EC	MA	MF	RP
Clean	95.3	91.2	91.4	89.3	94.1	91.9
10	94.1	88.1	89.6	85.7	89.4	90.2
5	91.6	78.7	78.6	74.2	87.7	81.8
0	79.8	67.5	66.7	60.5	76.1	68.5
Total	90.2	91.4	81.6	77.4	86.8	83.0

pared to those using audio queries with length of 3, 4, and 5 seconds.

Table 5 presents the results of the advertisement identification performed on set II database under five different noisy environments when a 5-second-long query was used. Compared with the identification results in Table 3, the identification results in Table 5 are not higher, because some TV advertisements contain silent segments and their fingerprints were random frequently used to query the fingerprint database for the retrieval process. As shown in Table 5, the proposed method, PM, achieves the best identification rates. The reason is that non-repeating foreground audio is extracted from repeating background audio effectively and applied to the generation of peak-pair audio fingerprints.

Table 6 presents the results of the music identification performed on the set I database under various types of distortions: resampling (RS), equalization (EQ), noise addition (NA), echo addition (EA), MP3 encoding/decoding (MP), pitch-shifting (PS), and time-stretching (TS).

As shown in Table 6, the proposed method, PM, yields overall the highest identification rate compared to other methods. The next best method is MF, which provides an especially better identification rate than WL, EC, MA, and RP. The identification rates of EC and RP are very similar. Their identification rates are better than those of WL and MA when the query samples suffer from a linear speed change (time-stretching) or the pitch of the query samples is changed without affecting its speed (pitch-shifting).

Compared with the identification results in Tables 6 and 7 presents the results of the advertisement identification performed on the set II database under various types of distortions.

Table 6 Comparative performance of six schemes with set I

Distortion	AIAR using 5-second-long query (%)					
	PM	WL	EC	MA	MF	RP
RS	98.3	93.1	94.5	92.2	96.1	95.1
EQ	98.3	92.8	93.8	92.2	96.1	95.1
NA	93.2	83.5	84.1	80.0	89.4	85.9
EA	92.9	82.2	83.9	79.3	89.1	84.6
MP	97.8	92.5	93.8	91.6	95.3	94.7
PS	91.4	23.7	55.3	37.8	80.6	52.3
TS	90.5	22.5	42.3	36.5	81.3	41.7
Total	94.6	70.0	78.2	72.8	89.7	78.5

Table 7 Comparative performance of six schemes with set II

Distortion	AIAR using 5-second-long query (%)					
	PM	WL	EC	MA	MF	RP
RS	95.3	91.2	91.4	89.3	94.1	91.9
EQ	95.3	91.2	91.4	89.3	94.1	91.9
NA	90.2	81.4	81.6	77.4	86.8	83.0
EA	89.5	80.7	79.6	76.2	87.8	82.7
MP	94.7	90.8	90.6	87.8	92.7	90.5
PS	88.2	20.3	52.5	35.2	77.8	49.8
TS	87.5	19.3	39.7	33.7	79.5	37.6
Total	91.5	67.8	75.3	69.8	87.5	75.3

The proposed PM has respectable performance results compared to WL, EC, MA, MF, and RP. From the experimental results in Tables 6 and 7, we believe that the significant improvement in various types of distortions is mainly due to the fact that the proposed hash function based on the peak-pair of non-repeating foreground audio has the effect to handle time-stretching, pitch-shifting, compression, and other modification as resampling and equalization.

4 Conclusion

In this paper, we proposed an audio fingerprinting method that can deal with various distortions by using a spectrogram of the modulated complex lapped transform, a extraction of non-repeating foreground audio from repeating background audio, an adaptive thresholding method for prominent peak detection as well as effective matching. Experimental results show that the proposed algorithm enhances Wang's fingerprint algorithm significantly, and achieves better identification rates compared to other methods. It is rather robust to different audio distortions and audio variations. In addition, it is suitable, due to its low computational complexity, for many practical portable consumer devices.

Enhancement and optimization of the extended search algorithm can be considered in future work. To improve the system the MCLT could be replaced by an efficient implementation of the non-stationary Gabor transform. The method will be applied to content retrieval, audio-based indoor localization and security applications running on smart TVs and mobile phones.

Acknowledgments This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2013R1A1A2007601). And this research was supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2015-H8501-15-1016) supervised by the IITP (Institute for Information & communications Technology Promotion), and Kwangwoon Research Grant 2014.

References

1. Cano, P., Batlle, E., Kalker, T., Haitsma, J.: A review of algorithms for audio fingerprinting. In: International Workshop on Multimedia Signal Processing, pp. 169–173 (2002)
2. Li, W., Xiao, C., Liu, Y.: Low-order auditory Zernike moment: a novel approach for robust music identification in the compressed domain. EURASIP J. Adv. Sig. Process. **1**, 1–15 (2013)
3. Sinitsyn, A.: Duplicate song detection using audio fingerprinting for consumer electronics devices. In: IEEE International Symposium on Consumer Electronics (ISCE06), St. Petersburg, Russia, pp. 1–6 (2006)
4. Cerquides, J.: A real time audio fingerprinting system for advertisement tracking and reporting in FM radio. In: 17th International Conference on Radioelektronika, Brno, Czech, pp. 1–4 (2007)
5. Haitsma, J., Kalker, T.: A highly robust audio fingerprinting system. In: 3rd International Society for Music Information Retrieval Conference (ISMIR), Paris, France, pp. 107–115 (2002)
6. Liu, Y., Yun, H.-S., Kim, N.S.: Audio fingerprinting based on multiple hashing in DCT domain. IEEE Sig. Process. Lett. **6**(6), 525–528 (2009)
7. Chandrasekhar, V., Sharifi, M., Ross, D.A.: Survey and evaluation of audio fingerprinting schemes for mobile query-by-example applications. In: 12th International Society for Music Information Retrieval Conference (ISMIR), Miami, USA, pp. 801–806 (2011)
8. Pan, X., Yu, X., Deng, J., Yang, W., Wang, H.: Audio fingerprinting based on local energy centroid. In: IET International Communication Conference on Wireless Mobile and Computing (CCWMC), Shanghai, China, pp. 351–354 (2011)
9. Baluja, S., Covel, M.: Audio fingerprinting: combining computer vision and data-stream processing. In: International Conference on Acoustic, Speech, and Signal Processing (ICASSP), Honolulu, Hawaii, pp. 2:213–2:216 (2007)
10. Anguera, X., Garzon, A., Adamek, T.: MASK: robust local feature for audio fingerprinting. In: International Conference on Multimedia and Expo (ICME), pp. 455–460 (2012)
11. Wang, A.: An industrial strength audio search algorithm. In: 4th International Society for Music Information Retrieval Conference (ISMIR), Baltimore, pp. 7–13 (2003)
12. Kim, H.-G., Kim, J.Y.: Robust audio fingerprinting method using prominent peak pair based on modulated complex lapped transform. ETRI J. **36**(6), 999–1007 (2014)
13. Fenet, S., Richard, G., Grenier, Y.: A scalable audio fingerprint method with robustness to pitch-shifting. In: 12th International Society for Music Information Retrieval Conference, Taipei, Taiwan, pp. 121–126 (2011)
14. Malvar, H.: Fast algorithm for the modulated complex lapped transform. IEEE Sig. Process. Lett. **10**(1), 8–10 (2003)
15. Rafii, Z., Pardo, B.: Repeating pattern extraction technique (REPET): a simple method for music/voice separation. EEE Trans. Audio Speech Lang. Process. **21**(1), 73–84 (2013)
16. Liutkus, A., Rafii, Z., Badeau, R., Pardo, B., Richard, G.: Adaptive filtering for music/voice separation exploiting the repeating musical structure. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, pp. 53–56 (2012)

Hyoung-Gook Kim received a Dr.-Ing. degree in Electrical Engineering and Computer Science from the Technical University of Berlin, Berlin, Germany. From 1998 to 2005, he worked on mobile service robots at Daimler Benz, and speech recognition at Siemens, Berlin, Germany. From 2005 to 2007, he was a project leader at the Samsung Advanced Institute of Technology, Korea. Since 2007, he has been a professor in the Department of Electronics Convergence Engineering, Kwangwoon University, Seoul, Korea. His research interests include audio signal processing, audiovisual content indexing and retrieval, and speech enhancement.

Hye-Seung Cho received the B.S. degree in Electronics Convergence Engineering from the Kwangwoon University, Seoul, Korea. Since 2015, she has been in the M.S. program in the Department of Radio Sciences and Engineering from Kwangwoon University, Seoul, Korea. Her research interests are audio signal processing, audiovisual content indexing and retrieval, and source separation.

Jin Young Kim received the Ph.D. degree in Electronic Engineering from the Seoul National University, Seoul, Korea. He worked on speech synthesis at Korea Telecom from 1993 to 1994. Since 1995 he has been a professor in the Department of Electronics and Computer Engineering, Chonnam National University, Gwangju, Korea. His research interests are speech synthesis, speech and speaker recognition, and audio-visual speech processing.